基于机器学习模型的海河北系干旱预测研究

赵美言[1]， 胡　涛[1]， 张玉虎[2]， 蒲　晓[2]， 高　峰[3]

（1　首都师范大学数学科学学院，北京　100048；2　首都师范大学资源环境与旅游学院，北京　100048；
3　国家气象信息中心，北京　100081 ）

摘　要： 提高干旱预测精度能为流域干旱应对及风险防范提供可靠数据支撑，构建比选合适的干旱模型是当前研究的热点。研究以4个时间尺度（3、6、9、12月）标准化降水指数（*SPI*）为表征指标，利用小波神经网络（WNN）、支持向量回归（SVR）、随机森林（RF）三种机器学习算法分别构建了海河北系干旱预测模型，利用Kendall、K–S、*MAE*三种检验方法判定模型表现及其稳定性。研究表明：（1）WNN、SVR模型呈现结果在不同时间尺度*SPI*存在差异，WNN最适合12个月尺度*SPI*干旱预测；SVR最适合6个月尺度*SPI*干旱预测。（2）对3、12个月尺度*SPI*，RF预测性能最优（Kendall > 0.898，*MAE* < 0.05）；对6、9个月尺度*SPI*，SVR预测性能最优（Kendall > 0.95，*MAE* < 0.04）。（3）模型预测性能稳定性存在区别，RF预测稳定性最高，其次为SVR。（4）构建的三种模型表现异同主要是因为SVR转为凸优化问题解决了WNN易陷入局部最优解的不足，从而提高了模型预测性能，RF集成多样化回归树，降低了弱学习器的负面影响，提高了模型预测准确率及稳定性，同时，RF处理包含噪声的降水数据的能力更强。

关 键 词： 干旱；WNN；SVR；RF；*SPI*；海河北系

文章编号： 1000－6060(2020)04－0880－09(0880～0888)

干旱是最常见、最复杂、对人类社会影响最为严重的气象灾害之一[1]，随着气候变暖海河流域干旱严重程度趋于上升且发生范围越来越广[2-3]，由干旱引起的旱灾程度在不断加重。提早开展干旱预测预报能够及时建立干旱预警机制，进行有效防范，减少干旱对人民生命财产及生态环境的影响。因此，如何提高干旱预测准确性、可靠性，建立干旱预测模型及遴选合适的模型工具是急需研究探讨的热点问题。

目前，国内外常用于干旱预测的方法有马尔科夫链[4-6]、灰色系统[7-8]、差分自回归移动平均[9-10]等。机器学习模型因其强大的预测能力[11-12]，在干旱预测领域也得到广泛应用[9,13-14]。常见的机器学习模型有小波神经网络（WNN）、支持向量回归（SVR）、随机森林（RF）、人工神经网络（ANN）等。WNN作为小波变换和ANN的结合，具有优于ANN的非线性处理能力，被大量用于干旱预测研究，如ZHANG[15]利用6、12个月尺度标准化降水指数（*SPI*），对海河北系进行实证预测，证实WNN优于ANN的拟合能力。支持向量回归（SVR）由SVM分类问题扩展而来，采用结构风险最小化的设计，适用于小样本数据、非线性问题以及高维数据[16-17]。SVR也得到了广泛应用，如Aminnejad[18]使用*SPI*和SVR对乌米亚湖盆地干旱进行预测，预测准确率在75%以上；措姆[13]采用SVR，以3、6、9个月尺度*SPI*作为研究对象，预测流域尺度的气象干旱，说明了SVR预测精度优于数据处理组合方法，二者均证实了SVR模型在干旱预测领域的适用性。尽管WNN、SVR已被证明可以用于干旱预测，但WNN和SVR模型也存在着预测稳定性不强，受

收稿日期： 2019－11－12； **修订日期：** 2020－03－19
基金项目： 北京市科技计划课题（编号：Z201100006720001）；首都师范大学交叉研究院项目（编号：00719530011010, 00719530012012, 00719530012010）；国家重点研发计划项目（编号：2017YFC0406002）
作者简介： 赵美言（1996－），女，硕士研究生，吉林省磐石市人，主要从事统计学在水文气象中应用. E-mail：zmy2180502132@163.com
通迅作者： 张玉虎（1975－），男，博士，副教授，主要从事环境系统分析及风险评估研究. E-mail：zhang_yuhu@163.com

*SPI*时间尺度影响较大等问题[15,19]。RF是一种基于分类回归树的组合模型，具有稳定的预测性能，同时可以处理包含噪声的预测变量，在预测研究中表现较好[20]，吴晶[14]利用*SPI*进行干旱等级分类，用随机森林模型对淮河流域进行干旱预测，整体预测平均准确率73.0%；沈润平[21]基于综合气象干旱指数使用RF模型对河南省构建遥感干旱监测模型，监测值和实测综合气象干旱指数值干旱等级的一致率达到81%。以上案例说明了这几种机器模型能够开展干旱预测。

但是，前人开展干旱预测成果多是构建单一算法模型，以单个时间尺度干旱指标为研究对象[10,22-26]，缺少多模型多时间尺度的综合对比分析。利用RF模型与WNN、SVR模型在同一研究区，不同时间尺度，对比分析干旱预测效果的文献却鲜见。同时，前人文献大多没有对几种模型及其结果稳定性开展分析，更是缺少几种模型算法结果表现差异的内在统计机理的探讨分析。基于此，本研究借助*SPI*的4种时间尺度(3月、6月、9月、12月)值，构建评价了海河流域北系WNN、SVR、RF三种模型干旱预测表现及其稳定性，初步探讨了模型差异化的内在机理，并确定最优干旱模型。研究结果为该地区或其他地区开展干旱预测提供了有益尝试。

1 研究区概况

1.1 研究区概况

海河北系地处北京、天津的上游地区，主要包括蓟运河、潮白河、北运河、永定河等河流(图1)，是我国重要的工农业生产区。流域面积为8.34×10[4] km[2]，其中山区、平原分别占62.5%、37.5%，属温带东亚季风气候，多年平均降水量约490 mm。近年来，海河北系降水整体偏枯[27]，海河流域干旱程度及干旱范围均呈上升趋势[2,28]，仅1961—2011年海河流域干旱发生次数达48次以上[29]。

1.2 数据

本文所选用数据来源于中国气象局(http://www.cma.gov.cn/)，选取了海河北系8个国家基准气象站点(表1)1960—2010年的逐日降水数据，并对数据进行了严格的修订和质量控制，降水缺失数据取附近平均值替代，确保数据的采集时间连续、完整。本文基于1960—2010年日降水数据，计算得到609个3个月尺度*SPI*、606个6个月尺度*SPI*、603个9个

表1 气象站点信息

Tab. 1 Information of the meteorological stations

站名	所属省份	经度/°E	纬度/°N	高程/m	平均年降水/mm	最大年降水/mm
北京	北京	116.5	39.8	31.3	370.2	579.0
大同	山西	113.3	40.1	1 067.2	398.9	616.3
丰宁	河北	116.6	41.2	661.2	457.9	696.4
怀来	河北	115.5	40.4	536.8	399.0	591.5
唐山	河北	118.2	39.7	27.8	379.0	543.6
蔚县	河北	114.6	39.8	909.5	711.9	1 193.4
张家口	河北	114.9	40.8	724.2	549.2	913.2
遵化	河北	118.0	40.2	54.9	605.4	1 007.7

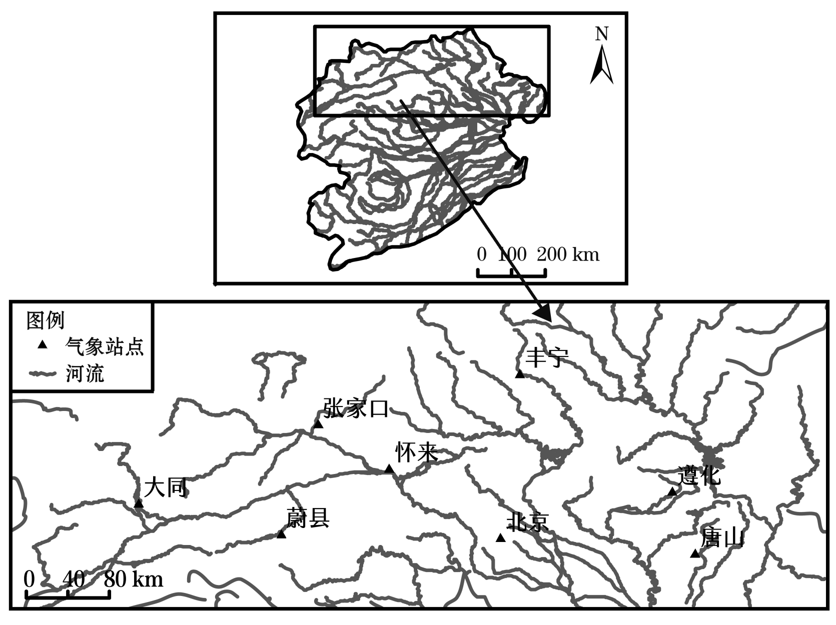


图1 海河北系气象站点分布

Fig. 1 Distribution of meteorological station of the north of the northern part of Haihe river basin

月的尺度 *SPI*、600个12个月尺度 *SPI*。*SPI* 计算方法请见参考文献30。

2 模型算法

2.1 小波神经网络(WNN)

WNN在传统的BP神经网络模型中融合了小波变换理论,是一种新型前馈神经网络模型,WNN使用小波函数作为BP神经网络隐含层神经元的激发函数[31]。记模型输入量为 x_i (i=1,…,k),输入层与隐藏层的连接权重为 ω_{ij},隐藏层与输出层的连接权重为 ω_{jk},小波基函数为 h_j,h_j 的平移因子为 b_j,h_j 的伸缩因子为 a_j,则隐藏层神经元的输出为:

$$(hj)=h_j\left[\frac{\sum_{i=1}^{k}\omega_{ij}x_i-b_j}{a_j}\right],\ j=1,\cdots,l \tag{1}$$

式中:l 为隐藏层神经元数目。记隐藏层第 i 个神经元输出结果为 $h(i)$,m 为输出层神经元数目,则输出层神经元输出结果为:

$$y(k)=\sum_{j=1}^{l}\omega_{jk}h(i),k=1,\cdots,m \tag{2}$$

本文使用的WNN包括输入层、隐藏层和输出层三层,使用Morlet母小波基函数。Morlet母小波基函数公式如下:

$$y=\cos(1.75x)\mathrm{e}^{\frac{x^2}{2}} \tag{3}$$

WNN以最小化均方误差为原则,利用梯度下降算法逐步调整网络连接权值与小波基函数的平移因子、尺度因子,以使网络的预测输出不断逼近期望输出。WNN的训练过程分为以下步骤:

步骤1:设定学习率 η 与隐藏层神经元个数 l,随机化网络连接权重 ω_{ij}、ω_{jk} 及小波函数伸缩因子 a_j、平移因子 b_j,;步骤2:分割数据为训练集和测试集,使用训练集数据训练网络,使用测试集数据计算网络预测精度;步骤3:依次向网络输入训练样本,计算网络输出与相应的预测误差e,利用误差e的反向传播修正网络权值和小波函数参数;步骤4:判断算法是否结束,如没有结束,返回步骤3。

2.2 支持向量回归(SVR)

SVR作为SVM处理拟合回归问题的一类模型,通过建立训练数据中待预测向量与支持向量间的非线性关系,可以对测试数据的待预测向量进行预测[32]。SVR的基本原理:给定训练集样本 $D=\{(x_1,y_1),(x_2,y_2),\cdots(x_l,y_l)\}\subset\aleph\times R$,$\aleph$ 为输入模式的空间。引入 ε 不敏感函数作为损失函数:

$$L_\varepsilon\left(f(x_i)-y_i\right)=\begin{cases}0, & \left|f(x_i)-y_i\right|<\varepsilon\\ \left|f(x_i)-y_i\right|-\varepsilon, & \left|f(x_i)-y_i\right|\geqslant\varepsilon\end{cases} \tag{4}$$

SVR算法的目的是寻找使得函数 $f(x_i)=\omega x_i+b$ 尽可能逼近实测值的参数对(ω,b)。

引入松弛因子 ξ,ξ^*,根据统计学习理论的结构风险化准则,回归问题转化为求解如下凸规化问题:

$$\min_{\omega,\xi,\xi^*}\left(\frac{1}{2}\|\omega\|^2+\sum_{i=1}^{l}C\left(\xi_i+\xi_i^*\right)\right) \tag{5}$$

$$s.t.:\omega x_i+b-y_i\leqslant\varepsilon+\xi_i^*,y_i-\omega x_i-b\leqslant\varepsilon+\xi_i$$

其中 $\xi_i\geqslant0,\xi_i^*\geqslant0$。引入拉格朗日函数,同时通过相应的鞍点条件简化得到:

$$\begin{aligned}&\min_{\alpha,\alpha^*}\sum_{i=1}^{l}\alpha_i(\varepsilon-y_i)-\sum_{i=1}^{l}\alpha_i^*(\varepsilon+y_i)\\&+\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\left(\alpha_i^*\alpha_i^*-\alpha_j\alpha_j\right)x_i^Tx_j\\&s.t.:\sum_{i=1}^{l}\left(\alpha_i^*-\alpha_i\right)=0,0\leqslant\alpha_i^*,\alpha_i\leqslant C\end{aligned} \tag{6}$$

由于原始优化问题有不等式约束,需要满足如下KKT(Karush-Kuhn-tucher)条件:

$$\left.\begin{aligned}&\alpha_i^*\left(y_i-f(x_i)-\varepsilon-\xi_i^*\right)=0\\&\alpha_i\left(f(x_i)-y_i-\varepsilon-\xi_i\right)=0\\&\alpha_i^*\alpha_i=0;\xi_i\xi_i^*=0\\&\left(C-\alpha_i\right)\xi_i=0;\left(C-\alpha_i^*\right)\xi_i^*=0\end{aligned}\right\} \tag{7}$$

通过序列最小优化算法得到支持向量决策模型:

$$f(x)=\omega^Tx+b=\sum_{i=1}^{l}\left(\alpha_i-\alpha_i^*\right)x_i^Tx+b \tag{8}$$

进一步引入核函数,SVR可表示为:

$$f(x)=\sum_{i=1}^{l}\left(\alpha_i^*+\alpha_i\right)k(x_i,x)+b \tag{9}$$

2.3 随机森林(RF)

RF是由Leo Breiman[33-34]在2001年提出的一种统计学习理论,是基于分类回归树的组合模型,既可用以数据分类,又能处理回归问题。RF的基本思想是利用自助(bootstrap)重采样技术,从总体训练样本集S中有放回等概率地重复抽样生成 K 个新的训练样本集 $C_1^*,\cdots,C_K^*$,每个训练样本集对应一棵决策树。在每棵树的结点,随机选取若干个特征进行节点分裂,并按照节点不纯度最小原则选择一个特征对该节点进行分裂。每颗决策树都得到最大限

度的生长,不进行剪枝操作,最终形成一个多元非线性组合模型。对于新输入数据,回归模型使用所有决策树的预测平均值作为最终预测结果,分类模型的预测结果由投票法则决定。RF回归模型算法原理如图2所示。

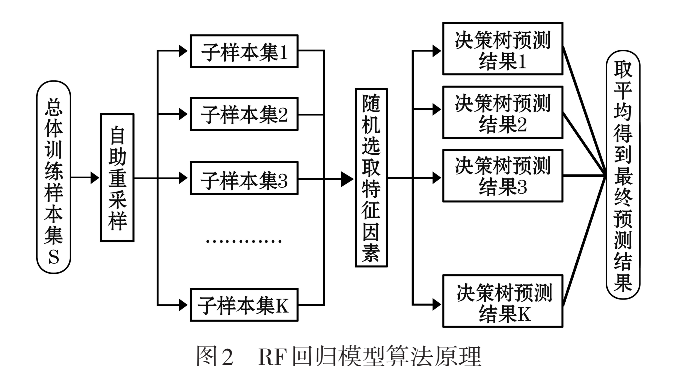


图2　RF回归模型算法原理

Fig. 2　Algorithm principle of RF regression model

3　模型构建与验证

3.1　模型构建

各模型在不同站点之间的建模过程相似,因此本文仅以北京站点为例介绍模型构建过程。模型构建仅使用*SPI*数据,以*SPI*历史数据作为模型输入变量,以当前*SPI*值作为输出变量。将1960年到2002年4月的*SPI*作为分析建模数据,2002年5月到2010年的*SPI*数据作为检验数据,对预测模型的有效性和稳定性进行检验。

WNN模型需要调节的超参数包括为输入层节点数、隐藏层节点数和学习率,输入层节点数即为延时阶数(预测当前*SPI*所需历史数据的个数)。构建WNN模型时,首先分割分析建模数据为85%的训练集和15%的验证集;而后使用训练集构建模型,使用验证集计算平均绝对误差(*MAE*),并利用网格搜索以*MAE*最小为原则寻找网络最优超参数,最后使用分析建模数据结合最优超参数构建WNN模型。表2为模型调参结果。

表2　WNN模型调参结果

Tab. 2　Results of WNN model parameter adjustment

尺度	学习率	输入层节点数	隐藏层节点数
SPI-3	0.01	6	12
SPI-6	0.001	2	18
SPI-9	0.001	2	6
SPI-12	0.001	2	6

SVR、RF模型的构建过程与WNN模型类似,仅需要优化的参数存在区别。在SVR、RF模型构建过程中,发现仅延时阶数对模型预测性能影响较大,本文所取延时阶数范围为1-15。图3、4显示RF、SVR模型的*MAE*均在延时阶数为3时达到最低值,据此得到,3为北京站点模型预测的最优延时阶数。

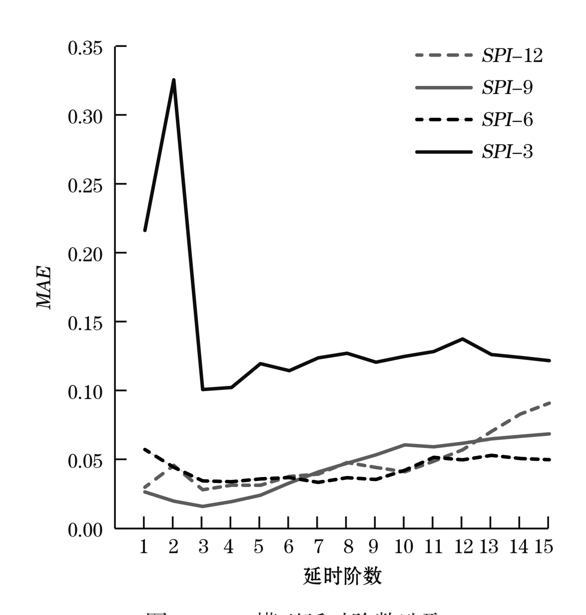


图3　SVR模型延时阶数选取

Fig. 3　Selection of lag order of SVR model

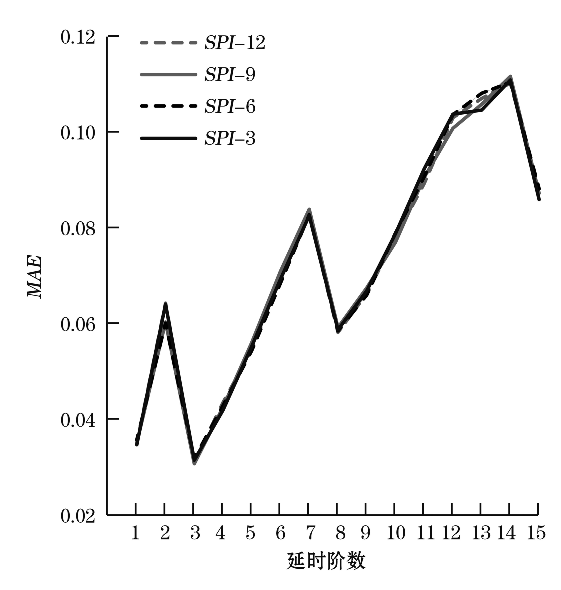


图4　RF模型延时阶数选取

Fig. 4　Selection of lag order of SVR model

3.2　模型评价

本文采用*MAE*、Kendall秩相关系数(Kendall)、Kolmogorov-Smirnov(K-S)检验定量评估模型预测表现。*MAE*用以描述两样本接近程度(预测值与实测值),值趋近0则说明两样本接近程度高;Kendall描

述预测值同实测值相关程度,值越接近1越好;K-S检验基于R语言实现,计算结果为两样本经验分布函数之间的绝对值最大距离,记为*D*统计量。*D*值越接近于0,两个样本来自相同分布的可能性越大。

3.3 模型检验

3.3.1 对比分析 对各研究站点3月尺度*SPI*建模预测,结果见表3。WNN的Kendall均低于0.5,K-S检验均高于0.2,*MAE*均高于0.4,说明WNN模型不能有效反映3个月尺度*SPI*的波动变化。SVR、RF的预测表现明显优于WNN,其Kendall达到0.85以上,K-S检验均在0.2之下。同时,RF的预测性能远高于SVR,其Kendall的平均值较SVR高约3.7%,K-S检验平均值较SVR低约30.9%,*MAE*平均值较SVR低约66.7%。

基于3月尺度*SPI*研究方法,得到各模型在6月、9月、12月*SPI*预测结果(表4、表5、表6)。对6个月尺度*SPI*,WNN模型预测表现最差,最优预测结果均在SVR模型出现,且其*MAE*平均值约为RF的50%,表明对于6个月尺度*SPI*,SVR预测性能最优。

对9月尺度*SPI*,SVR的Kendall最高,均接近1,且除站点丰宁外,其*MAE*最低,虽个别站点其K-S检验值高于RF,但均在0.1之下,综合所有评价指标,在9月尺度*SPI*上SVR预测性能最优。

对12月尺度*SPI*,WNN的*MAE*均在0.15以上;SVR的K-S检验在站点丰宁、怀来、遵化、北京高于0.1;RF在各评价指标值的表现均比较优异,其Kendall均高于0.9,K-S检验不高于0.1,*MAE*均低于0.04。综合表明,对12月尺度*SPI*,RF预测性能最优。

3.3.2 稳定性分析 通过计算评价指标的站点平均值,探究*SPI*时间尺度变化对模型预测性能稳定性的影响(见图5)。WNN的预测性能随着*SPI*时间尺度的变化表现出明显差异,其评价指标Kendall、K-S检验*MAE*平均值的极差分别为0.351、0.277、0.265,且对12月尺度*SPI*的预测性能最优,各评价指标值都显著改善。

SVR的预测性能随着*SPI*时间尺度的变化表现出轻微差异,其评价指标Kendall、K-S检验、MAE平均值的极差分别为0.079、0.064、0.067,且对6月尺度*SPI*预测性能最优。

表3 *SPI*-3序列各模型比较

Tab. 3 Comparison of models of *SPI*-3 sequence

模型	指标	大同	蔚县	丰宁	张家口	怀来	遵化	北京	唐山
WNN	Kendall	0.469	0.477	0.508	0.435	0.493	0.523	0.460	0.470
	K-S检验	0.462	0.264	0.300	0.352	0.400	0.275	0.330	0.264
	MAE	0.426	0.462	0.462	0.443	0.476	0.441	0.437	0.462
SVR	Kendall	0.908	0.923	0.899	0.861	0.859	0.897	0.907	0.928
	K-S检验	0.121	0.110	0.121	0.132	0.089	0.132	0.110	0.165
	MAE	0.068	0.054	0.088	0.087	0.086	0.108	0.111	0.095
RF	Kendall	**0.928**	**0.936**	**0.929**	**0.898**	**0.937**	**0.952**	**0.925**	**0.944**
	K-S检验	**0.076**	**0.098**	**0.087**	**0.098**	**0.087**	**0.065**	**0.087**	**0.076**
	MAE	**0.018**	**0.033**	**0.027**	**0.027**	**0.029**	**0.033**	**0.034**	**0.035**

表4 *SPI*-6序列各模型比较

Tab. 4 Comparison of models of *SPI*-6 sequence

模型	指标	大同	蔚县	丰宁	张家口	怀来	遵化	北京	唐山
WNN	Kendall	0.712	0.716	0.691	0.708	0.699	0.700	0.711	0.710
	K-S检验	0.173	0.136	0.163	0.199	0.176	0.156	0.154	0.181
	MAE	0.328	0.346	0.392	0.358	0.356	0.383	0.316	0.382
SVR	Kendall	**0.979**	**0.985**	**0.979**	**0.979**	**0.976**	**0.980**	**0.973**	**0.975**
	K-S检验	**0.044**	**0.033**	**0.055**	**0.055**	**0.088**	**0.055**	**0.077**	**0.066**
	MAE	**0.012**	**0.027**	**0.012**	**0.017**	**0.023**	**0.024**	**0.033**	**0.015**
RF	Kendall	0.960	0.966	0.956	0.931	0.953	0.950	0.962	0.945
	K-S检验	0.055	0.088	0.077	0.077	**0.088**	0.066	0.044	0.077
	MAE	0.024	0.033	0.030	0.032	0.037	0.036	**0.033**	0.035

表5 *SPI*-9序列各模型比较

Tab. 5 Comparison of models of *SPI*-9 sequence

模型	指标	大同	蔚县	丰宁	张家口	怀来	遵化	北京	唐山
WNN	Kendall	0.762	0.727	0.628	0.584	0.642	0.702	0.726	0.676
	K-S检验	0.176	0.143	0.198	0.154	0.132	0.264	0.132	0.176
	MAE	0.277	0.238	0.260	0.284	0.262	0.273	0.251	0.281
SVR	Kendall	**0.981**	**0.993**	**0.995**	**0.979**	**0.998**	**0.991**	**0.980**	**0.993**
	K-S检验	**0.044**	0.088	0.132	0.066	**0.088**	0.099	**0.044**	0.077
	MAE	**0.020**	**0.028**	0.036	**0.027**	**0.017**	**0.026**	**0.016**	**0.022**
RF	Kendall	0.973	0.958	0.942	0.935	0.946	0.940	0.953	0.950
	K-S检验	0.066	**0.055**	**0.088**	**0.055**	**0.055**	**0.088**	**0.044**	**0.055**
	MAE	0.028	0.032	**0.029**	0.036	0.029	0.031	0.030	0.033

表6 *SPI*-12序列各模型比较

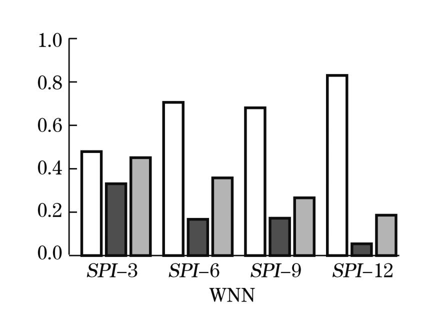
Tab. 6 Comparison of models of *SPI*-12 sequence

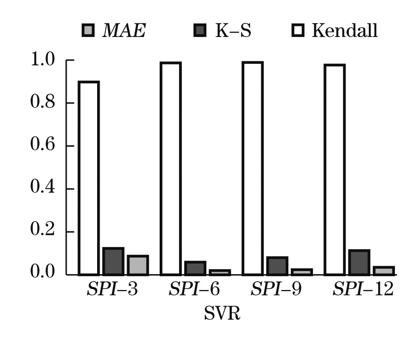
模型	指标	大同	蔚县	丰宁	张家口	怀来	遵化	北京	唐山
WNN	Kendall	0.820	0.833	0.816	0.824	0.838	0.865	0.826	0.816
	K-S检验	**0.047**	0.044	**0.057**	**0.035**	**0.057**	**0.052**	0.077	0.063
	MAE	0.229	0.173	0.151	0.193	0.178	0.164	0.207	0.192
SVR	Kendall	**0.961**	0.917	**0.998**	**0.979**	**0.996**	**0.989**	**0.987**	**0.992**
	K-S检验	0.067	0.090	0.222	0.067	0.111	0.167	0.111	0.067
	MAE	0.049	0.064	0.039	**0.025**	**0.024**	**0.032**	**0.028**	**0.020**
RF	Kendall	0.958	**0.949**	0.936	0.955	0.941	0.948	0.957	0.948
	K-S检验	0.067	**0.033**	0.078	0.078	0.067	0.100	**0.056**	**0.044**
	MAE	**0.035**	**0.029**	**0.026**	0.029	0.026	0.033	0.031	0.027

RF的预测性能在不同时间尺度*SPI*的表现无明显差异,其评价指标Kendall、K-S检验、MAE平均值的极差分别为0.003、0.022、0.021;综上,三种模型中,WNN对*SPI*时间尺度的变化最为敏感,模型的预测性能最不稳定;RF对*SPI*时间尺度的变化最不敏感,模型的预测性能最稳定。

4 讨论

研究基于1960–2010年日降水数据,以*SPI*作为干旱指标,利用WNN、SVR、RF三种模型分别开展海河北系干旱预测,利用Kendall、K-S检验、MAE分别评价了模型预测结果表现。SVR、RF模型预测性能优于WNN,二者能够准确反映各时间尺度*SPI*序列的波动变化。WNN模型预测性能最差。SVR、RF具有不同时间尺度*SPI*的适用性,对3、12月尺度*SPI*,RF预测性能优于SVR,而对另两个时间尺度,SVR优于RF。尽管SVR在个别时间尺度*SPI*的预测性能优于RF,但RF的预测性能稳定性强于





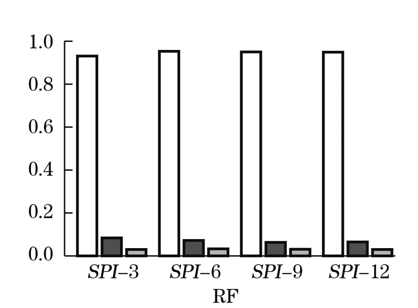


图5 模型稳定性比较

Fig. 5 Comparison of model stability

SVR。3种预测模型中,WNN调参过程最为复杂且模型运算速度最慢,SVR、RF的调参过程相较简单,仅需优化滞后阶数,且运算速度快。模型预测性能的差异主要由以下方面导致:(1)WNN模型基于ANN模型,存在收敛到局部最优解的问题,模型预测准确度因此得不到保证。(2)SVR模型通过转化为凸优化问题,避免了WNN模型陷入局部最优解的问题,因此提升了模型预测精度。(3)RF模型为一种集成学习算法,多样化的回归树能够有效提高弱学习器的预测效果,从而提高了模型预测的准确率与稳定性,且RF对噪声具有较好的容忍性[35],在处理含有噪声的降水数据时优势更大。鉴于各模型预测表现,开展干旱预测或预警分析时,建议灵活选用模型。对本文已探讨过的时间尺度的*SPI*,依据性能最优选择预测模型,对多时间尺度*SPI*进行的预测研究,建议选择预测性能优异且预测稳定性最强的RF模型。未来可以进一步比较探究SVR、RF在更长时间尺度以及其他地区干旱预测的适用性,同时,三种模型预测结果轨迹不同的内在统计机制也需更进一步研究。

5 结 论

(1) 三种机器学习模型在不同时间尺度*SPI*预测表现分别为,WNN最适用于12个月尺度*SPI*的预测;SVR最适用于6个月尺度*SPI*的预测;RF对不同时间尺度*SPI*预测效果无明显区别。

(2) 在同一时间尺度上,对于3、12个月尺度*SPI*,RF具有最优的预测性能(Kendall > 0.898,*MAE* < 0.05),能够较准确反映*SPI*真值的变化情况;对于6、9个月尺度*SPI*,SVR具有最优的预测性能(Kendall > 0.95,*MAE* < 0.04),且对于6个月尺度*SPI*,SVR模型的预测性能为本文所有预测研究中最优的。

(3) 从*SPI*时间尺度变化对模型预测性能影响的角度来看,WNN模型的稳定性最差,RF预测性能稳定性最高,其评价指标Kendall、K–S检验、*MAE*的平均值在不同时间尺度*SPI*极差最低,分别为0.003、0.022、0.021。

参考文献(References)

[1] 高涛涛,殷淑燕,王水霞. 基于SPEI指数的秦岭南北地区干旱时空变化特征[J]. 干旱区地理,2018,41(4):85-94.[GAO Taotao, YIN Shuyan, WANG Shuixia. Spatial and temporal variations of drought in northern and southern regions of Qinling Mountains based on standardized precipitation evapotranspiration index[J]. Arid Land Geography,2018,41(4):85-94.]

[2] 王文静,延军平,刘永林,等. 基于综合气象干旱指数的海河流域干旱特征分析[J]. 干旱区地理,2016,39(2):334-336.[WANG Wenjing,YAN Junping,LIU Yonglin,et al. Characteristics of droughts in the Haihe Basin based on meteorological drought composite index[J]. Arid Land Geography,2016,39(2):334-336.]

[3] 倪深海,顾颖,彭岳津. 近七十年中国干旱灾害时空格局及演变[J]. 自然灾害学报,2019,28(6):176-181.[MI Haishen,GU Yin,PENG Yuejin. Patio-temporal pattern and evolution trend of drought disaster in China in recent seventy years[J]. Journal of Natural Disasters,2019,28(6):176-181.]

[4] ZHU S, LUO X, CHEN S, et al. Improved hidden markov model incorporated with Copula for probabilistic seasonal drought forecasting [J]. Journal of Hydrologic Engineering, 2020,25(6).

[5] 王志成. 基于改进马尔柯夫链的区域干旱预测[J]. 水资源开发与管理,2018,(2):55-57.[Wang Zhicheng. Regional drought prediction based on improved Markov chain[J]. Water Resources Development and Management,2018,(2):55-57.]

[6] 马齐云,张继权,王永芳,等. 内蒙古牧区牧草生长季干旱特征及预测研究[J]. 干旱区资源与环境,2016,30(7):157-163.[MA Qiyun,ZHANG Jiquan,WANG Yongfang,et al. Characteristics and prediction of drought in growing season in Inner Mongolia pastoral area[J]. Journal of Arid Land Resources and Environment,2016,30(7):157-163.]

[7] 韩会明,刘喆玥,刘成林,等. 灰色模型的改进及其在气象干旱预测中的应用[J]. 南水北调与水利科技,2019,17(6):62-68.[HAN Huiming,LIU Zheyue,LIU Chenglin,et al. Improvement of grey model and its application in forecast of meteorological drought[J]. South-to-North Water Transfers and Water Science & Technology,2019,17(6):62-68.]

[8] 谷洪波,刘芷妤. 湖南农业旱灾的时间规律分析及重灾年份预测[J]. 湖南科技大学学报(社会科学版),2016,19(5):110-116.[GU Hongbo,LIU Zhiyu. Time regularity analysis and trend prediction of agricultural drought disaster in Hunan Province[J]. Journal of Hunan University of Science & Technology (Social Science Edition),2016,19(5):110-116.]

[9] 杨慧荣,张玉虎,崔恒建,等. ARIMA和ANN模型的干旱预测适用性研究[J]. 干旱区地理,2018,41(5):47-55.[YANG Huirong, ZHANG Yuhu, CUI Hengjian, et al. Applicability of ARIMA and ANN models for drought forecasting[J]. Arid Land Geography,2018,41(5):47-55.]

[10] ZHANG Y, YANG H, CUI H, et al. Comparison of the ability of ARIMA, WNN and SVM models for drought forecasting in the Sanjiang Plain, china [J]. Natural Resources Research, 2019,29:1447-1464.

[11] 杨海民,潘志松,白玮. 时间序列预测方法综述[J]. 计算机科学,2019,46(1):21-28.[YANG Haimin, PAN Zhisong, BAI Wei. A survey of time series prediction methods[J]. Computer

Science,2019,46(1):21-28.]

[12] 疏杏胜,王子茹,李福威. 基于机器学习模型的短期降雨多模式集成预报[J]. 南水北调与水利科技,2020,18(1):42-50. [SHU Xingsheng, WANG Ziru, LI Fuwei. Short-term rainfall multi-mode integrated forecasting based on machine learning models[J]. South-to-North Water Transfers and Water Science & Technology,2020,18(1):42-50.]

[13] 措姆,加勇次成,红梅. 利用数据挖掘方法探索流域尺度气象干旱预报的研究[J]. 四川环境,2018,37(4):65-70. [CUO Mu, JIAYONG Cicheng, HONG Mei. Using data mining methods to explore meteorological drought forecasts at river basin scales[J]. Sichuan Environment,2018,37(4):65-70.]

[14] 吴晶,陈元芳,余胜男. 基于随机森林模型的干旱预测研究[J]. 中国农村水利水电,2016,(11):17-22. [WU Jing, CHEN Yuanfang, YU Shengnan. Research on drought prediction based on random forest model[J]. China Rural Water and Hydropower,2016,(11):17-22.]

[15] ZHANG Y, LI W, CHEN Q, et al. Multi-models for *SPI* drought forecasting in the north of Haihe River Basin, China [J]. Stochastic Environmental Research & Risk Assessment, 2017,31(10):2471-2481.

[16] 张佼,田琦,王美萍. 基于交叉验证支持向量回归的供热负荷预测[J]. 中北大学学报(自然科学版),2014,35(5):189-206. [ZHANG Jiao, TIAN Qi, WANG Meiping. Heating load prediction for heating systems based on support vector regression with cross validation[J]. Journal of North University of China(Natural Science Edition),2014,35(5):189-206.]

[17] 王金安,李飞. 复杂地应力场反演优化算法及研究新进展[J]. 中国矿业大学学报,2015,44(2):189-205. [WANG Jinan, LI Fei. Review of inverse optimal algorithm of in-situ stress filed and new achievement[J]. Journal of China University of Mining & Technology,2015,44(2):189-205.]

[18] AHMADEBRAHIMPOUR E, AMINNEJAD B, KHALILI K. Application of global precipitation dataset for drought monitoring and forecasting over the Lake Urmia Basin with the GA-SVR model[J]. International Journal of Water, 2018, 12(3):262-277.

[19] 葛强. 基于随机森林的奎屯河水资源可持续利用评价[J]. 人民珠江,2019,40(1):79-83. [GE Qiang. Evaluation of sustainable utilization of water resources in Kuitun River based on random forest[J]. Pearl River,2019,40(1):79-83.]

[20] TYRALIS H, PAPACHARALAMPOUS G, LANGOUSIS A. A brief review of random forests for water scientists and practitioners and their recent history in water resources[J]. Water, 2019, 11(5):910.

[21] 沈润平,郭佳,张婧娴,等. 基于随机森林的遥感干旱监测模型的构建[J]. 地球信息科学学报,2017,19(1):125-133. [SHEN Runping, GUO Jia, ZHANG Jingxian, et al. Construction of a drought monitoring model using the random forest based remote sensing[J]. Journal of Geo-information Science, 2017, 19(1):125-133.]

[22] 张玉虎,向柳,孙庆,等. 贝叶斯框架的copula季节水文干旱预报模型构建及应用[J]. 地理科学,2016,36(9):1437-1444. [ZHANG Yuhu, LIU Xiang, SUN Qing, et al. Bayesian probabilistic forecasting of seasonal hydrological drought based on Copula function[J]. Scientia Geographica Sinica, 2016, 36(9): 1437-1444.]

[23] CAI W, ZHANG Y, YAO Y, et al. Probabilistic analysis of drought spatiotemporal characteristics in the Beijing-Tianjin-Hebei metropolitan area in China[J]. Atmosphere, 2015, 6(4): 431-450.

[24] ZHANG Y, YAO Y, LIN Y, et al. Satellite characterization of terrestrial drought over Xinjiang Uygur Autonomous Region of China over past three decades [J]. Environmental Earth Sciences,2016,75(6):451.

[25] ZHANG Y, XIE P, PU X, et al. Spatial and temporal variability of drought and precipitation using cluster analysis in Xinjiang, northwest China [J]. Asia Pacific Journal of Atmospheric Sciences,2019,55:155-164.

[26] ZHANG Y, CAI W, CHEN Q, et al. Analysis of changes in precipitation and drought in Aksu River Basin, northwest China [J]. Advances in Meteorology,2015,2015:1-15.

[27] 章数语,王建华,翟家齐. 海河北系1956年-2012年降水时序演变特征[J]. 南水北调与水利科技,2016,14(3):36-42. [ZHANG Shuyu, WANG Jianhua, ZHAI Jiaqi. Characteristics analysis of time serial of rainfall in the northern part of Haihe River Basin from 1956 to 2012[J]. South-to-North Water Transfers and Water Science & Technology,2016,14(3):36-42.]

[28] 宗燕,王艳君,翟建青. 海河流域气象干旱时空特征分析[J]. 干旱区资源与环境,2013,27(12):198-202. [ZONG Yan, WANG Yanjun, ZHAI Jianqing. Spatial and temporal characteristics of meteorological drought in the Haihe River Basin based on standardized precipitation index[J]. Journal of Arid Land Resources and Environment,2013,27(12):198-202.]

[29] HE J, YANG X, LI J, et al. Spatiotemporal variation of meteorological droughts based on the daily comprehensive drought index in the Haihe River Basin, China [J]. Natural Hazards, 2015,75(S-2):199-217.

[30] 李文卿,江源,赵守栋,等. 六盘山地区油松树轮宽度年表与多尺度标准化降水指数的关系[J]. 生态学报,2017,37(10): 3365-3374. [LI Wenqing, JIANG Yuan, ZHAO Shoudong, et al. Response of tree-ring width chronology of pinus tabulaeformis to multi-scale standardized precipitation index(*SPIn*) in the Liupan Mountain area[J]. Acta Ecologica Sinica, 2017, 37(10): 3365-3374.]

[31] 王宇,卢文喜,卞建民,等. 基于小波神经网络的地下水流数值模拟模型的替代模型研[J]. 中国环境科学,2015,35(1): 139-146. [WANG Yu, LU Wenxi, BIAN Jianmin, et al. Surrogate model of numerical simulation model of groundwater based on wavelet neural network[J]. China Environmental Science, 2015,35(1):139-146.]

[32] 王霞,王占岐,金贵,等. 基于核函数支持向量回归机的耕地面积预测[J]. 农业工程学报,2014,(4):204-211. [WANG Xia, WANG Zhanqi, JIN Gui, et al. Land reserve prediction using different kernel based support vector regression[J]. Transactions of the Chinese Society of Agricultural Engineering, 2014, (4):

204-211.]

[33] BREIMAN L. Random forests[J]. Machine Learning, 2001, 45(1): 5-32.

[34] BREIMAN L. Statistical modeling: The two cultures (with comments and a rejoinder by the author)[J]. Statistical Science, 2001, 16(3): 199-231.

[35] 王奕森,夏树涛. 集成学习之随机森[J]. 信息通信技术,2018,12(1):49-55.[WANG Yisen, XIA Shutao. A survey of random forests algorithms[J]. Information and Communications Technologies, 2018, 12(1): 49-55.]

Drought prediction based on machine learning models in the northern part of Haihe River Basin

ZHAO Mei-yan[1], HU Tao[1], ZHANG Yu-hu[2], PU Xiao[2], GAO Feng[3]

(1 *School of Mathematical Sciences, Capital Normal University, Beijing* 100048, *China*;
2 *College of Resources Environment & Tourism, Capital Normal University, Beijing* 100048, *China*;
3 *National Meteorological Information Center, Beijing* 100081, *China*)

Abstract: Drought is one of the major natural disasters. Improving the accuracy of drought prediction can provide reliable data to support drought response and risk prevention. The construction of suitable drought prediction models is a current research hotspot. Machine learning models are widely used for drought forecasting such as artificial neural network (ANN), wavelet neural network (WNN), support vector regression (SVR) and random forest (RF). This paper explored and compared the forecasting abilities and stabilities of the wavelet neural network (WNN), support vector regression (SVR) and random forest (RF) in the northern part of the Haihe River Basin, China. The northern part of the Haihe River Basin is located in the upper reaches of Beijing and Tianjin, which is an important industrial and agricultural production area in China. The total area is 8.34×10^5 km^2. It has a temperate monsoon climate with average annual precipitation of 490 mm. The models used in this paper are based on the standard precipitation index (*SPI*) at different time scales (3, 6, 9 and 12 months). The *SPI* was calculated using daily precipitation data obtained at eight meteorological points in the northern part of the Haihe River Basin from 1960 to 2010. Then, the *SPI* series were predicted use the WNN, SVR and RF models separately. The effectiveness of the three machine learning models is compared by Kendall rank correlation (Kendall), Kolmogorov-Smirnov (K-S) test and mean absolute error (*MAE*). The following results were observed: (1) The prediction abilities of the WNN and SVR models vary at different time scales, with WNN performing best suited for *SPI*-12 and SVR best suited for *SPI*-6. (2) For the *SPI*-3 and *SPI*-12, the RF prediction performance was optimal (Kendall > 0.898, $MAE < 0.05$). For the *SPI*-6 and *SPI*-9, the SVR prediction performance was optimal (Kendall > 0.95, $MAE < 0.04$). (3) The stability of the model prediction performances differed, with RF being most stable, followed by SVR. (4) The variation in model predictions performance is due to the following: the convex optimization of SVR resolves the WNN weakness of falling into a local optimal solution, thereby improving the prediction performance of the model. The RF boosting diversified regression trees, which reduce the negative influence of weak learners, improve the prediction accuracy and stability of the model. Furthermore, the capacity of the RF model is strongest in its ability to cope with precipitation data that contains noise. This paper presents a comprehensive analysis of the drought prediction performance of multiple models at multiple time scales for *SPI* series and preliminarily explores the internal mechanisms of model differentiation. The result of this study provides alternative models and research ideas for the northern part of the Haihe River Basin and beyond.

Key words: drought; SVR; RF; WNN; *SPI*; the northern part of Haihe River Basin